Analyze Google Analytics Data and Clustering - Predicting Users

Hanh Thi Lai

June – 2020

1. Introduction

1.1. Background

Google Analytics is a powerful tool where people can get many useful information about their potential customers, when they begin to visit until they leave website. People usualy using Google Analytics to view information at specific time to find more reasons to explain about how the revenue increase or decrease. The revenue increase because more people come to website, but how about the features of these people, what things we need to do in the future is sometimes not clearly. Therefore, it is advantageous for company to clustering customers to have more approriate sale – marketing campaigns and predict the users come to website to predict the revenue.

1.2. Problem

Data needed to analysis might include the date, page user visit, type of user (new user or return user), user's time on page, how many sessions, where user come from (google search or google adwords,...), which device they use, longitude and lattitude of user and how many goals. This project aims to clustering user and predict how many users will go to website in the future.

1.3. Interest

The company have set up Google Analytics on their website would be concern and interested in examines Google Analytics webiste data, how data can be turned into information, and how that information relates to a business or organization in the future. The information is very useful and also more authentic, more clearly than just guess.

2. Data acquisition and cleaning

2.1. Data sources

I get data by configuring and initialize my client's API with Python. After that, I used Python to extract the report from Google Analytics by defined the dimensions and metrics:

- Dimensions: ga:sourceMedium, ga:deviceCategory, ga:PagePath, ga:longitude, ga:latitude, ga:dateHour, ga:daysSinceLastSession
- Metrics: ga:users, ga:newUsers, ga:sessions, ga:goal1Completions, ga:timeOnPage, ga:pageviewsPerSession

2.2.Data cleaning and feature selection

Because of restriction of Google Analytics API so I have to get data in groups. I get total 3.340.291 entries and the result .csv file can not view in Notepad or Excel.

Data need be cleaning and preprocessing before could be used to analyze and apply machine learning models. First, I drop 404 pagePath because these error page is useless. After that, I create the group column extraced from the first part of PagePath to group data to category of products, so that category column can be used to analyze how users information differences in each product category. After that, I drop error rows, group rows into same category,...

Then I split sourceMedium to separate columns so it is more easily to analyze source information. I also extract search data from PagePath, so that I can create wordcloud for better visualize how frequency keywords users using in website. And I make some neccesary processing relate to null value, data type, not valid data... to make data is ready for using after that.

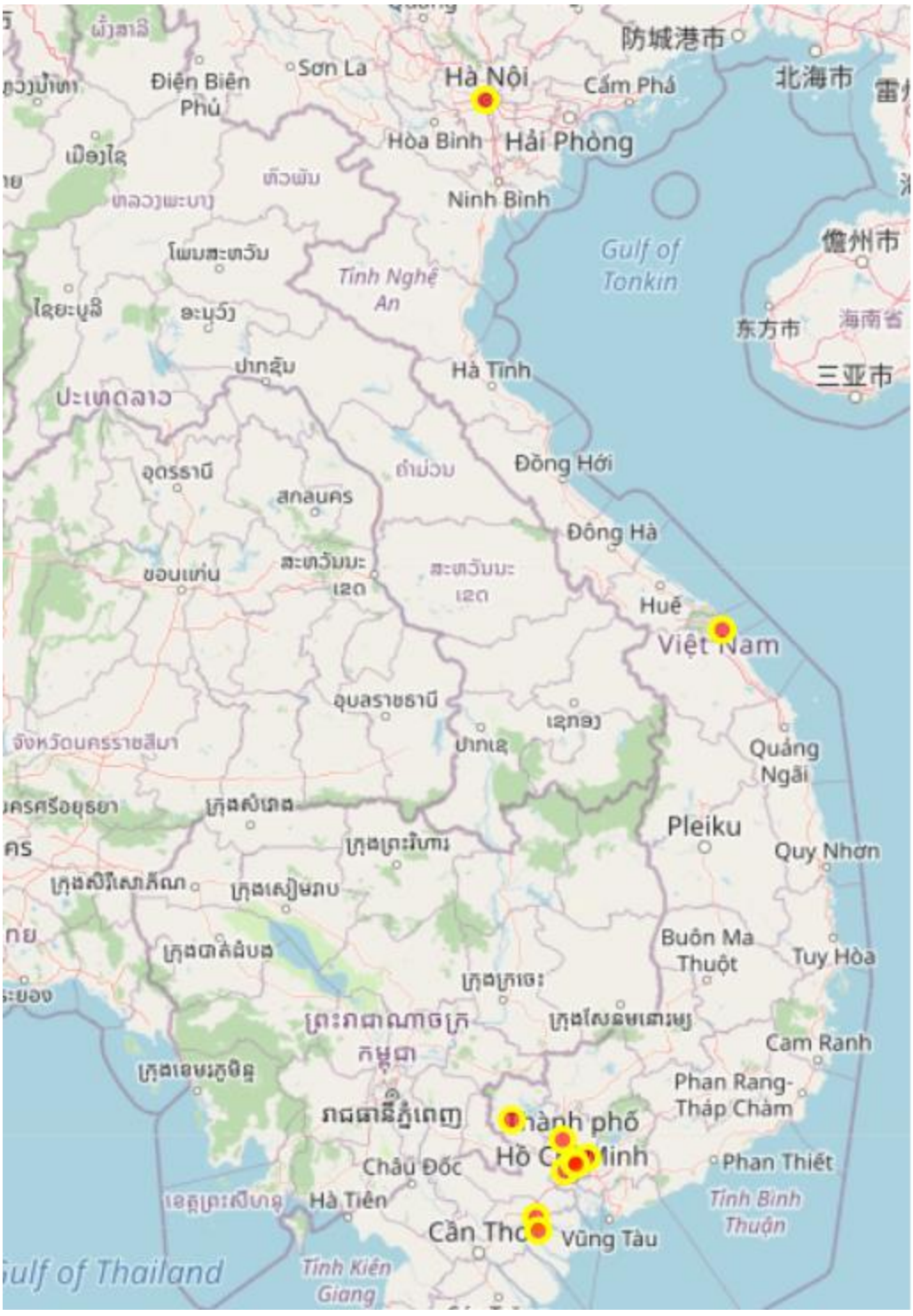
To predict in time series, I also copy data to another dataframe with only information needed to apply time series prediction (date, users, goals).

After data cleaning and preprocessing, there were 1.533.138 samples and 15 features in the data. This data is selected when collect so data is ready for exploratory and apply machine learning models.

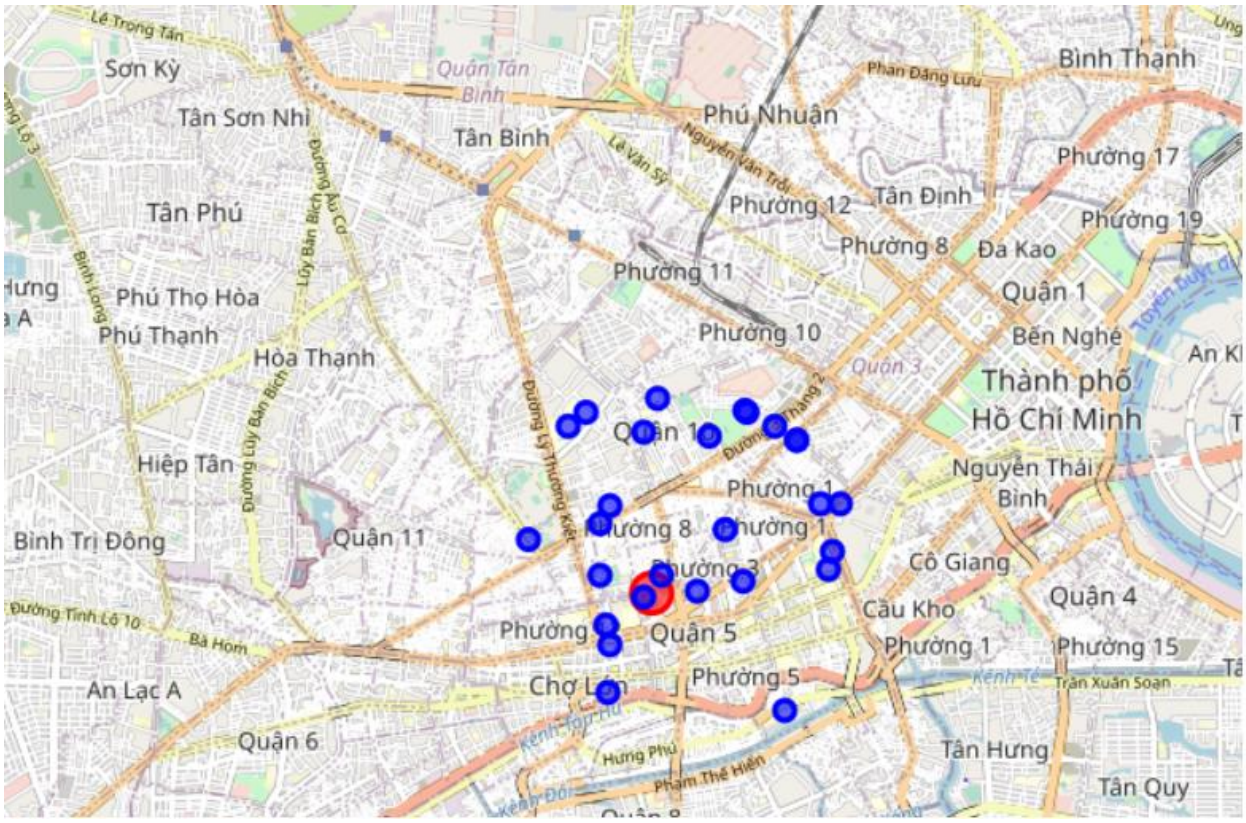
No	Column name	Data type	Meaning
1	device	object	Which device user using when connect to website
2	longitude	float64	Longitude of the user's city
3	latitude	float64	Latitude of the user's city
4	date	datetime64[ns]	Date when user visit website

5	day_of_week	int64	Day of the week, 0-Monday and 6 - Sunday
6	hour	Object	Hour (0-23) when user visit website
7	group	Object	Category of product
8	source	Object	Where user come from (google,ask, facebook,...)
9	medium	Object	Medium user come from (google, direct,...)
10	users	int64	Number of users
11	sessions	int64	Number of sessions
12	goals	int64	Number of goals
13	newUsers	int64	Number of new users
14	timeOnPage	float64	Average time user spend on page
15	pageviewsPerSession	float64	Average page views per session
16	daysSinceLastSession	float64	Average days since last session user spend on page

I used python folium library to visualize geographic details of top 10 venues have large number of users come into website. I used latitude and longitude values collected from Google Analytics to get the visual as below:



I also used Foursquare API to get the university (category id ="4d4b7105d754a06372d81259") near the first place above, as shown in result:



3. Methodology

3.1.Exploratory Data Analysis

3.1.1. How about user’device

I wonder what is the most popular device user using when visit website. Although mobile is very common now but desktop is still the top device of user. This plot will confirm that website interface should be optimized for desktop and mobile, over 90% users and new users.

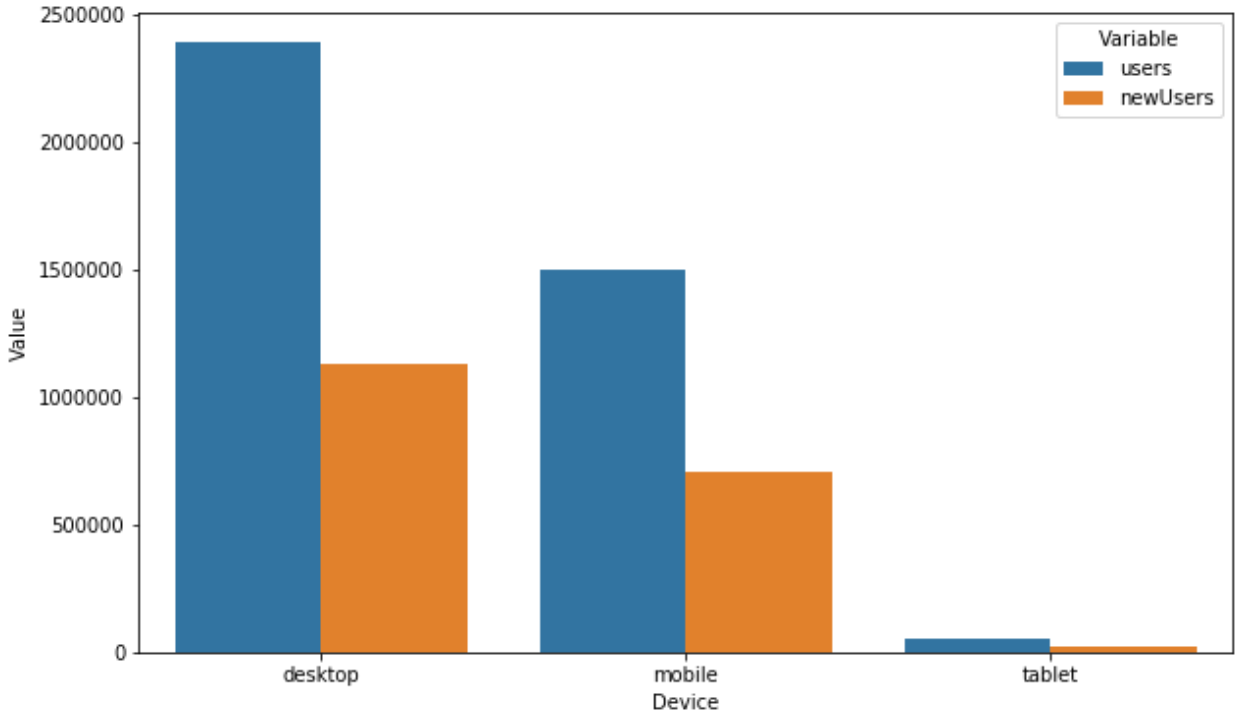


Figure 1. Total users and new Users among their device

Another important question is how about device user using when make a goal, register for a course. The answer is the mobile is same as desktop. So, the company should concern the process of register – check out – payment on mobile, besides the desktop.

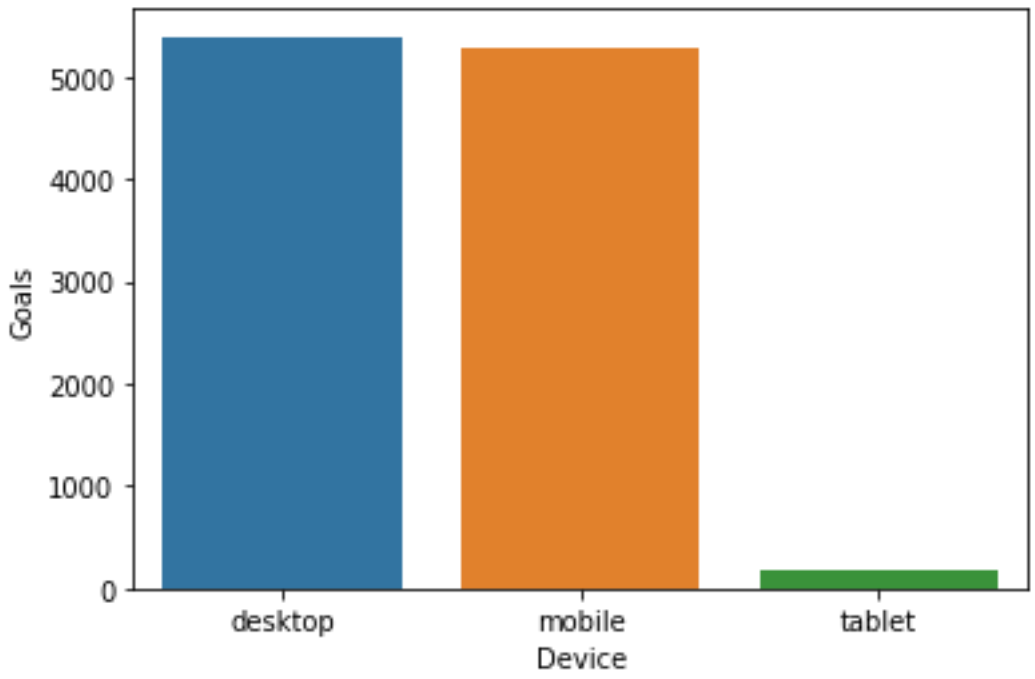


Figure 2. Total goals among their device

3.1.2. How about the category of courses

We have about 200 course and devide to 3 main categories: tin-hoc-van-phong, lap-trinh-va-csdl and do-hoa-da-truyen-thong. The website also contain some other pages grouped into home-page, chức-năng, khác and học viên. The following plot show how many users come to each catetory of content in our website.

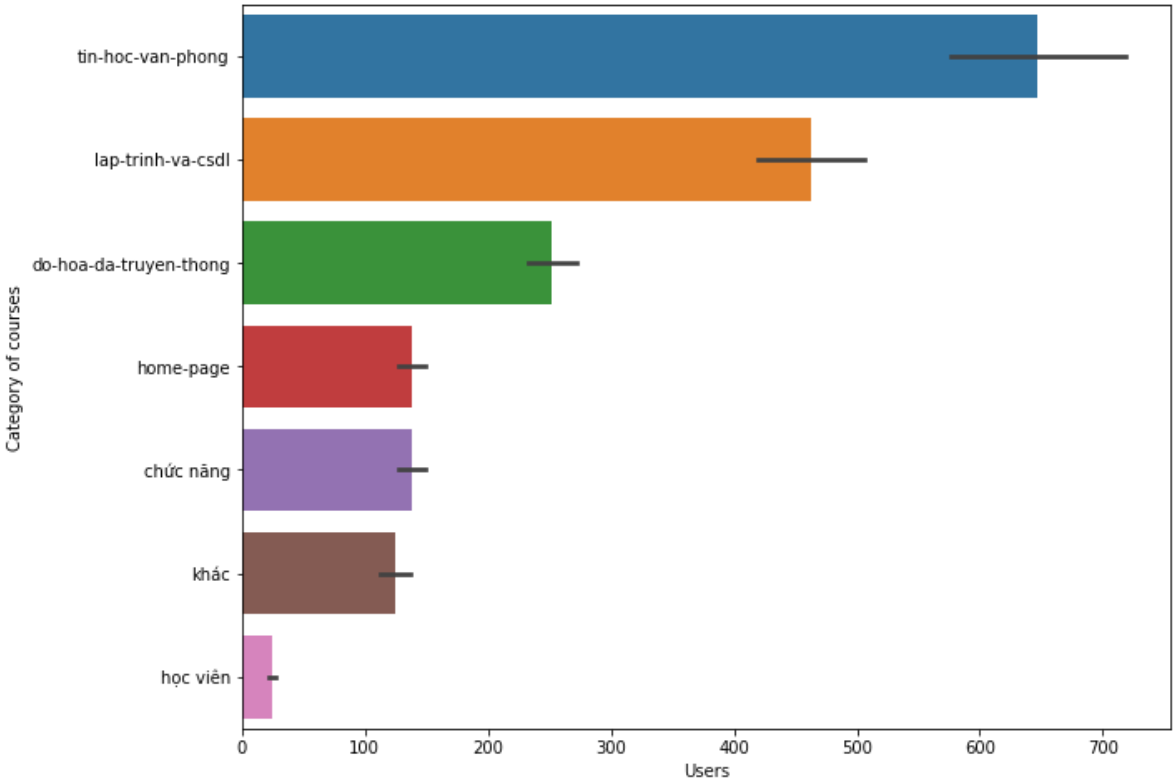


Figure 3. Total users among the course categories

Then I play attention to 3 main categories of course, with the following plot, I found that the large number of tin-hoc-van-phong's users is come from organic, the percentage between users come from organic of tin-hoc-van-phong is the largest.

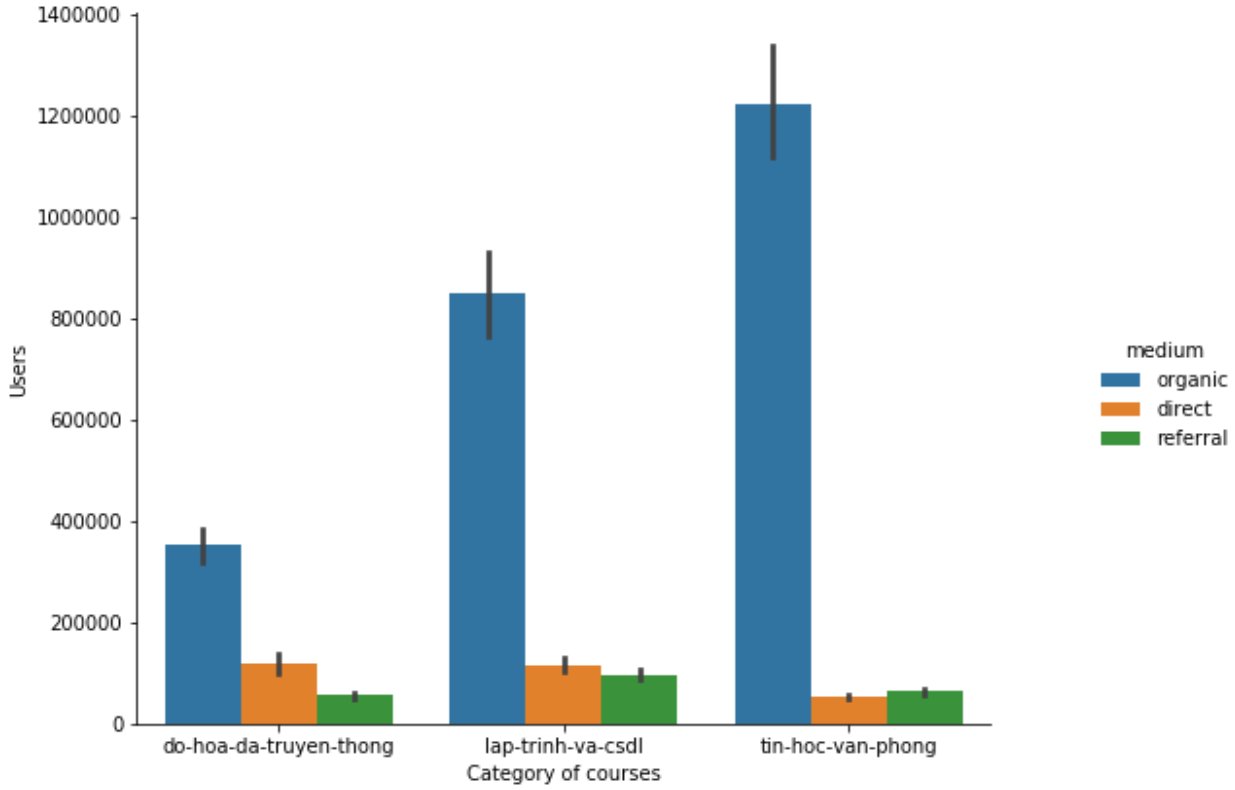


Figure 4. Total users among the course categories and source of user

When I analysis in number of goals (user register the course) in the fig... , I found that something interestings:

- The goals from lap-trinh-va-csdl come mainly from desktop and organic search.
- The goals from tin-hoc-van-phong come mainly from mobile and organic search.
- The goals from the direct and referral is smaller than organic search. So, company should concern on SEO to maintain and increase traffic from organic search.

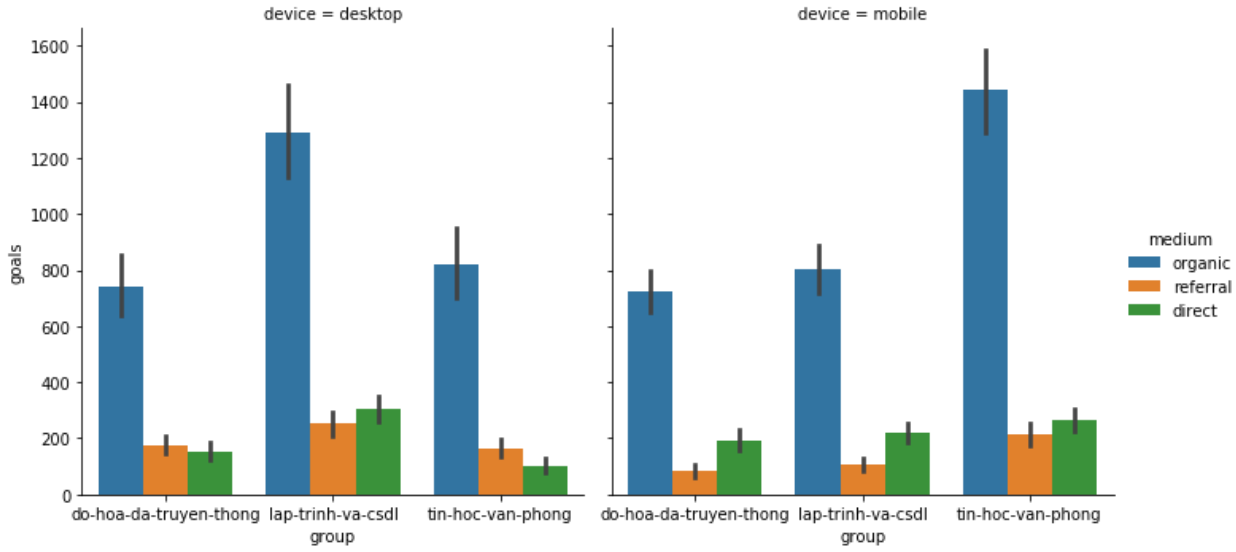


Figure 5. Total users among the course categories – medium - device

3.1.3. How about the relation between features

Firstly I visualize the relation between the number of users and number of goals, as the following plot:

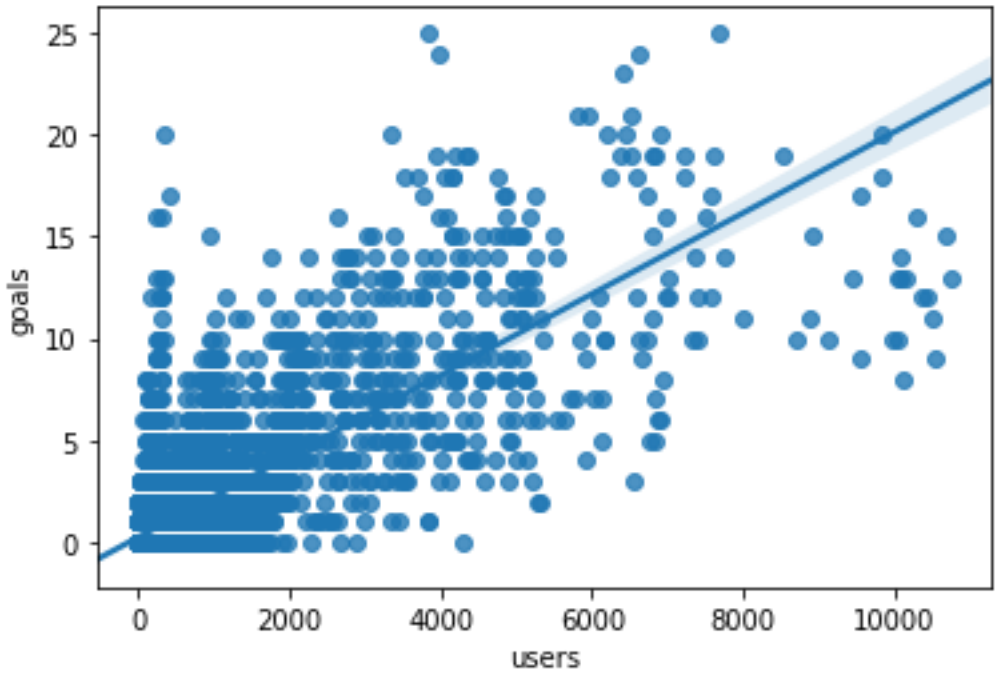


Figure 6. The relation between users and goals

When more users comming, the more they make the goals. So, marketing department should have more campaigns to get user to website, should pay attention to organic search, PR publications, optimize website for SEO, ...

Then I visualize the relation between timeOnPage and pageviewsPerSession, user usually pay more time when they view more Page in their session:

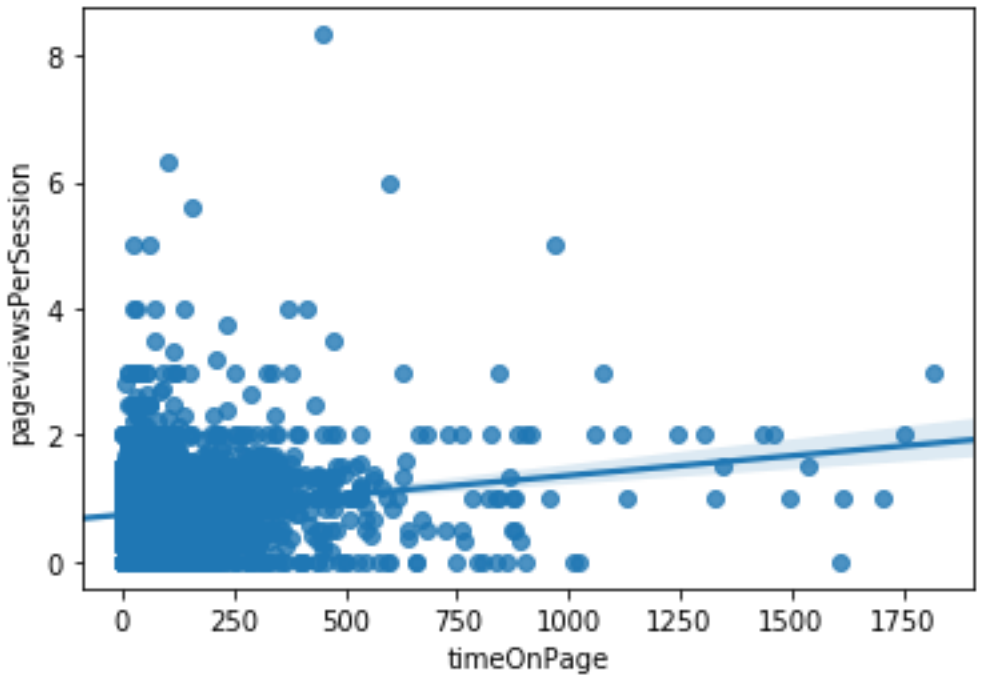


Figure 7. The relation between timeOnPage and pageviewsPerSession

3.1.4. How about the time

Then I plot the number of users come to website from 2017-2020, I see that the trend maybe stable in year 2020. Maybe COVID-19 makes the users somehow decrease in early months of 2020.

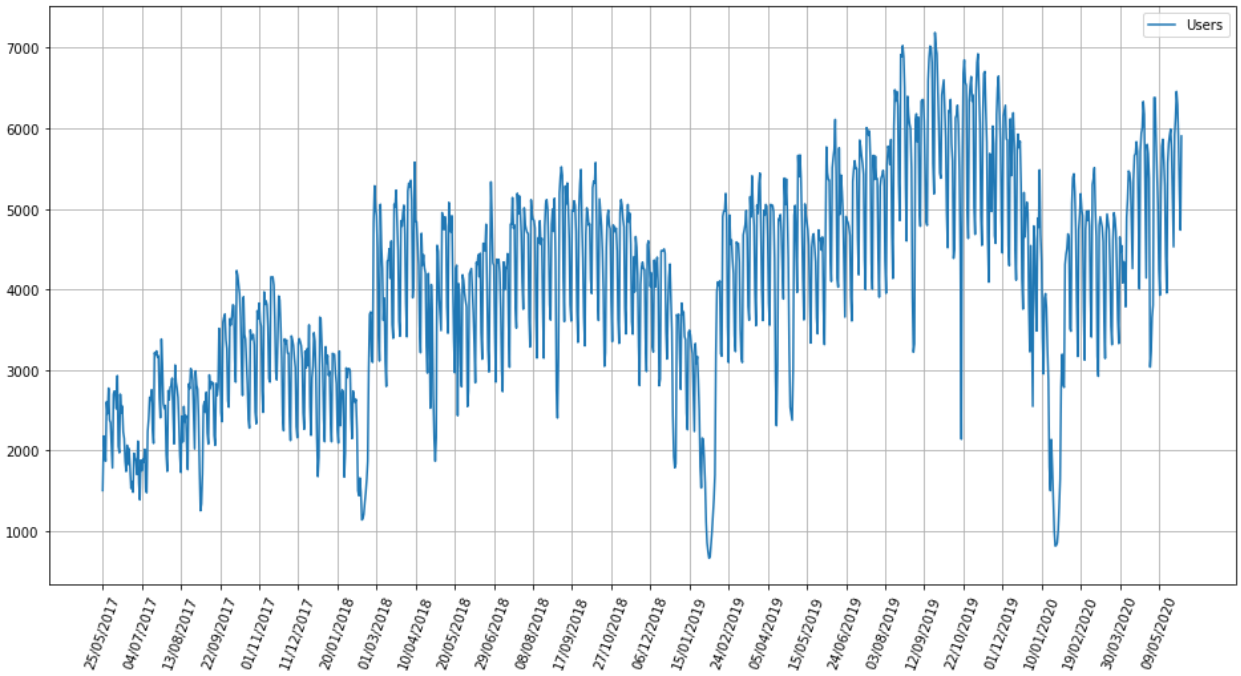


Figure 8. The trends of number users from 2017-2020

After that, I analize how change the number of users come to website in every month of the year, as in the following plot:

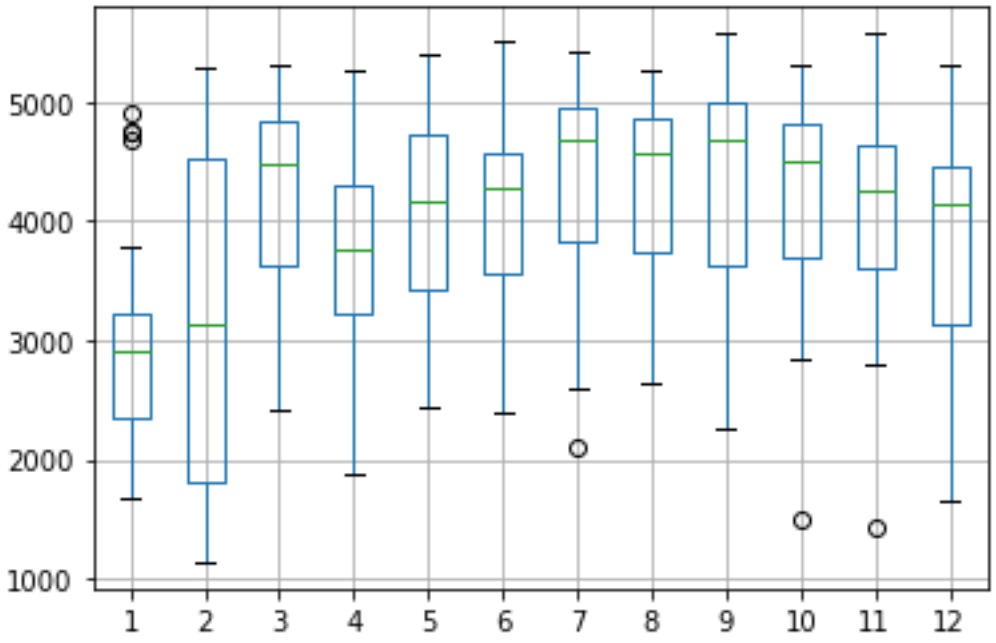


Figure 9. The number of users in each month of year

The users come to website at average about 4000 – 5000 users/month. But Janurary, April, November, December is not high as others month. February

is can not sure because it is Lunar New Year vacation (7-10 days off), but it is also the time user usually come to learn to begin a new year – new chance for future.

Then I visualize how number of users come to website change in the day of week and hour of day.

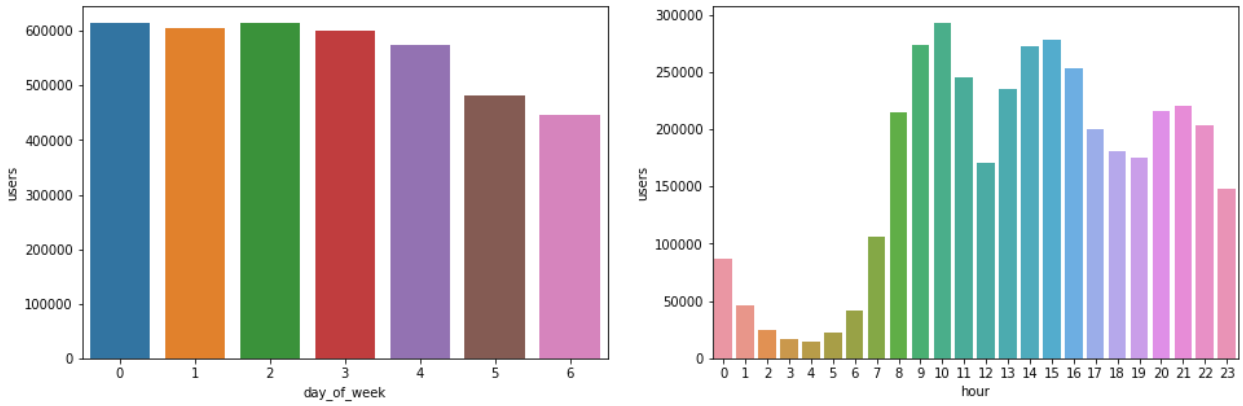


Figure 10. The number of users in each day of week and each hour of day

Users usually come to website on working day (from Monday to Friday) more than in weekend (Saturday and Sunday). And the time user usally come to website is after 8.am , same same as the goals:

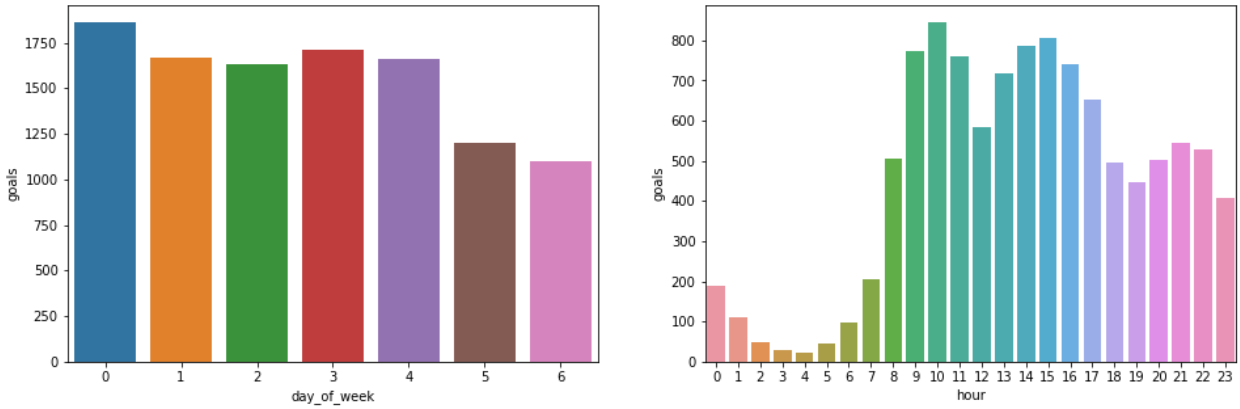


Figure 10. The number of goals in each day of week and each hour of day

3.2. Predictive Modeling

Clustering is an unsupervised machine learning technique to identify / group similar observations. Members of a group are very similar, and members of different groups are dissimilar.

In this section, I use are k-means clustering for creating customer segments based on following features:

- group_cat: based on group column, I transform group column to category data type and assign it with category code.
- medium_cat: based on medium column, I transform medium column to category data type and assign it with category code.
- device_cat: based on device column, I transform device column to category data type and assign it with category code.
- timeOnPage_cat: based on timeOnPage column, I transform to timeOnPage_cat by using pd.cut with bins=3
- pageviewsPerSession_cat: based on pageviewsPerSession column, I transform to pageviewsPerSession_cat by using pd.cut with bins=3
- daysSinceLastSession_cat: based on daysSinceLastSession column, I transform daysSinceLastSession_cat by using pd.cut with bins=3
- goals: number of goals
- users: number of users

When apply K-means clustering, the number of clusters K is predetermined. With algorithm iteratively I assigns each data point to one of the K clusters based on the feature similarity. As the number of clusters increase, the list variable ***sse*** keeps decreasing and then the rate of decrease slows down resulting in an elbow plot. The number of clusters at the elbow formation usually gives an indication on the optimum number of clusters.

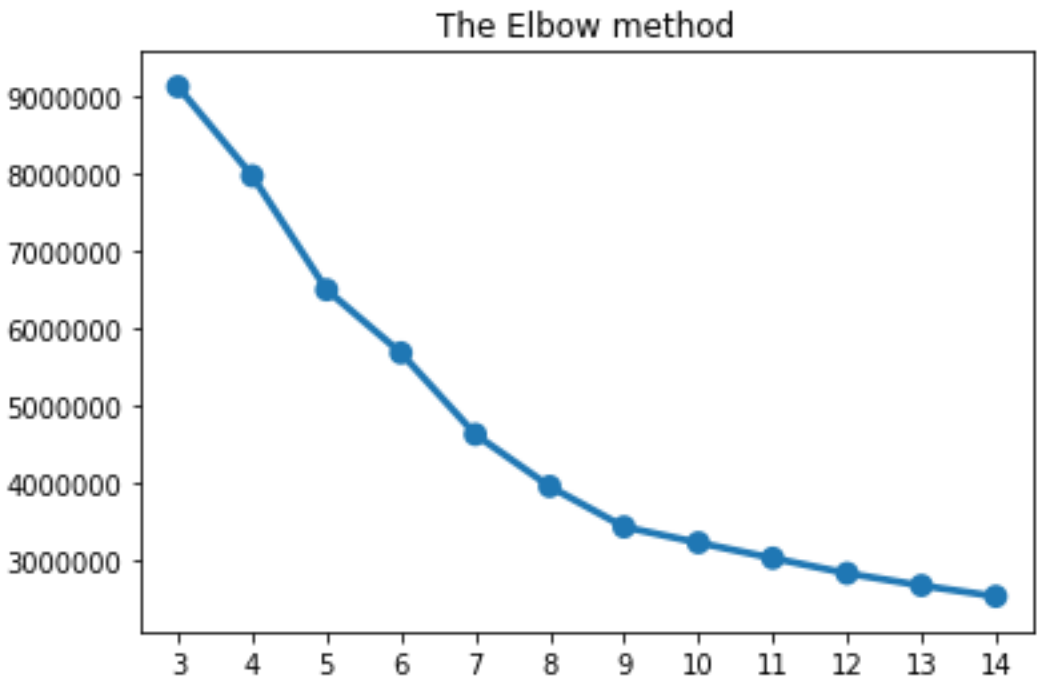


Figure 11. The elbow show the relation between number of cluster and sse error

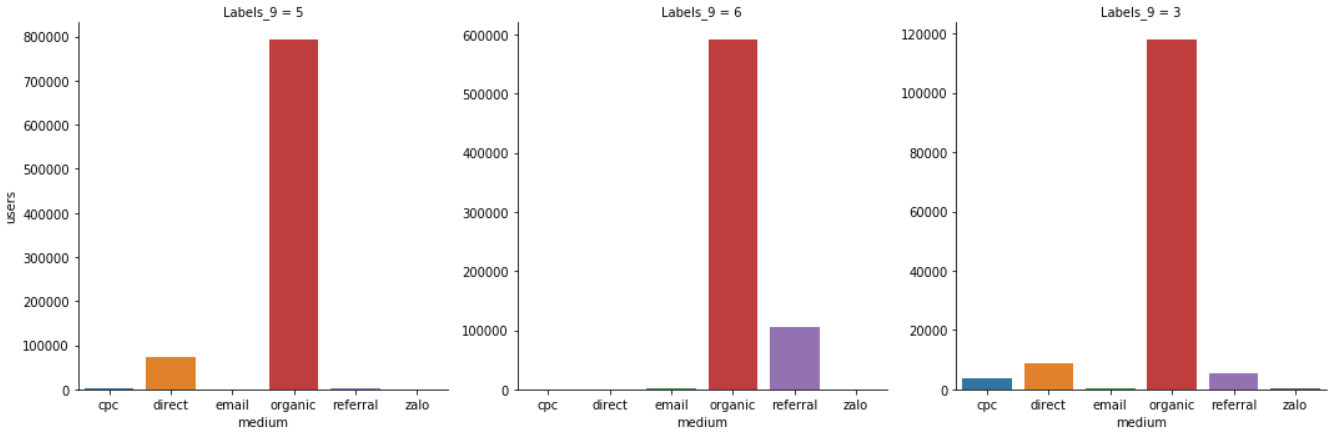
Based on that visualization, I will choose number of cluster equal to 9 with the information in each labels :

	timeOnPage	pageviewsPerSession	daysSinceLastSession	goals	users
	mean	mean	mean	sum	sum
Labels_9					
0	54.9	0.6	24.1	0	527820
1	32.4	0.8	0.1	0	497592
2	57.1	0.6	0.1	0	363335
3	113.4	0.6	4.5	10839	136401
4	7808.1	4.0	0.0	0	335
5	136.7	0.8	6.2	0	869537
6	29.8	0.9	0.1	0	696951
7	263.8	2.1	4.3	0	480770
8	51.3	0.7	2.2	0	365458

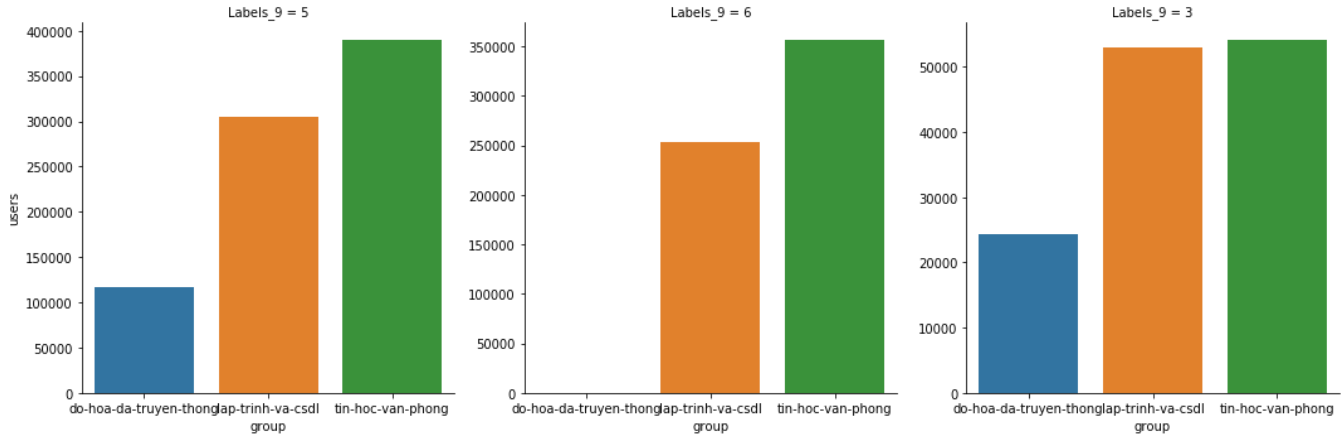
With the top number of users in cluster labels 5,6 and goals from cluster 3, I can explain the behavior of customer after clustering as follow:

- If user view one page in short time (average 30 seconds) they usually return to website immediately to view more information they need (Label =6)
- If user spend about 2 minutes on page, they usually return to website in about 6 days after last session (Label =5).
- User make goal will spend at average 113 seconds on page, view 0,6 pageview per session and return to website at average 4,5 days after last session. (Label =3)

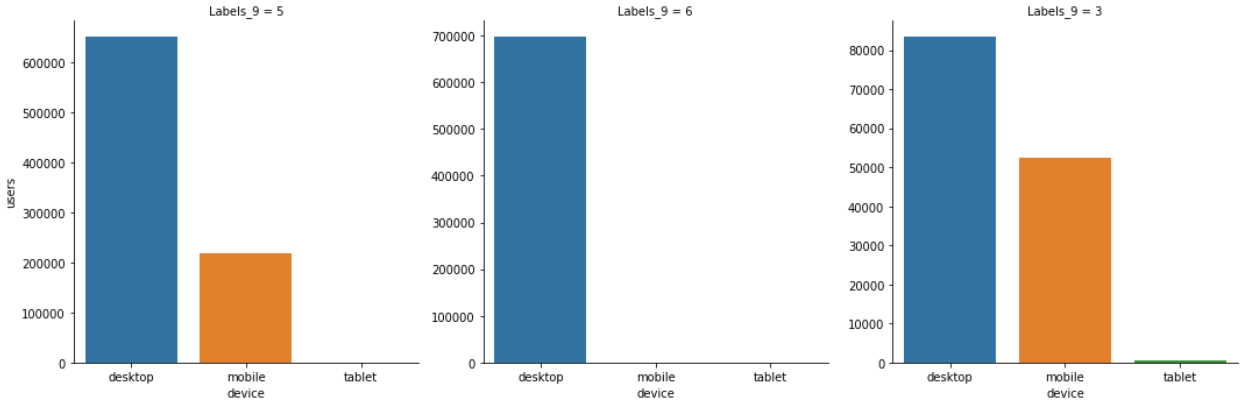
Users from Label =5 usually come from organic search and direct, but users from Label = 6 usually come from organic search and referral channel, user from Label=3 come from many source, as in the following plot:



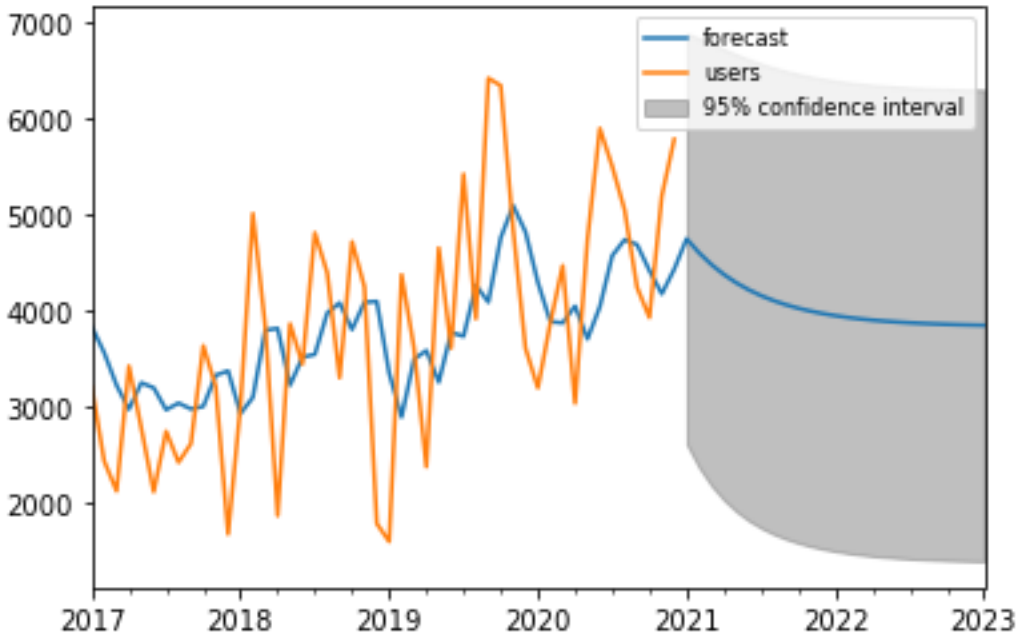
With users Label =5, there are 3 groups of category, but with Label=6 only tin-hoc-van-phong and lap-trinh-va-csdl group. Label=3 show that the number of goals of lap-trinh-va-csdl and tin-hoc-van-phong is the same.



With device, users in Label=5 use desktop and mobile but users in Label=6 only use desktop, user make goal from desktop more than using mobile (Label=3)



I also apply time series prediction to predict how about number of users come to website in near future, as in the following plot:



With this plot, the number of users come to website if do not have any improvements, the number of users in 2021 – 2023 will decrease.

4. Result and Discussion

My analysis so that some useful information can get from a large number of samples collected from Google Analytics:

- Although mobile is very common now but desktop is still the top device of users using when visit website.
- Number of user using mobile to register a course is same as number of user using desktop.
- The large number of tin-hoc-van-phong's users is come from organic, the percentage between users come from organic of tin-hoc-van-phong is the largest.
- The goals from lap-trinh-va-csdl come mainly from desktop and organic search.
- The goals from tin-hoc-van-phong come mainly from mobile and organic search.
- The goals from the direct and referral is smaller than organic search.
- When more users comming, the more they make the goals.
- Users usually come to website on working day (from Monday to Friday) more than in weekend (Saturday and Sunday). And the time user usally come to website is after 8.am , same same as the goals.

After apply k-means clustering to segmentation user, I can get useful insight:

- Label =6: User of tin-hoc-van-phong and lap-trinh-va-csdl group only use desktop come mainly from organic search and referral channel usualy view one page in short time (average 30 seconds) and return to website immediately to view more information they need.
- Label =5: If user spend about 2 minutes on page, they usually return to website in about 6 days after last session by using desktop or mobile, come from organic search and direct.
- Label =3: User make goal come mainly from organic, will spend at average 113 seconds on page, view 0,6 pageview per session and return to website at average 4,5 days after last session and the number of goals of lap-trinh-va-csdl and tin-hoc-van-phong is the same.

After apply time series prediction, I can get the information that the number of users in 2021 – 2023 will decrease.

I think there are some idea can make improvements for my company :

- Marketing department should have more campaigns to get user to website, should pay attention to organic search, PR publications, optimize website for SEO, ...
- Company should concern on SEO to maintain and increase traffic from organic search.
- Company should consider about budget on chanels: cpc, email, referral and should have KPI to check performance for these chanels.
- Website should have campaign to make user visit website more frequently because when they visit more likely they make goal.
- Mobile should develop more ease of use for user.
- Website should change content more frequently to get more user come in and return.
- If user spend more time on website, when they leave, should notice and get what information they concern, so that suggest them what they maybe need and they can more easily make a goal in the future.

5. Conclusion

In this study, I collect data from Google Analytics and implement user segmentation by using K-Means clustering based on category courses, medium,

device and mean of timeOnPage, pageviewsPerSession, daysSinceLastSession, number of goals, number of users. I also research and apply time series prediction in my study, it has very interesting result.

There are many problems when connect to Google API to get data, when processing data before apply machine learning model but the result is very interesting. It help me can propose some solutions for my company on how using budget more efficiently to get more users come to website. It also help any company using Google Analytics can get useful insight from that data.

I just use K-Means clustering with elbow method to choose the number of K cluster. However, if I can apply more clustering models, maybe I can evaluate which model is more suitable for analyze data from Google Analytics.

In addtion, another important information is to analyze the flow of user behavior from visit until to finish goal. It will help to increase the percentage user complete the goal, not just sum the number of goals. I will spend more time to improve the result in the future.